Link and Node Prediction in Metabolic Networks with Probabilistic Logic

Angelika Kimmig, Fabrizio Costa

{angelika.kimmig, fabrizio.costa}@cs.kuleuven.be
Department of Computer Science, Katholieke Universiteit Leuven, Belgium

We are nowadays capable of representing organism-wide metabolic processes. In fact there exist collections of metabolic networks for several hundreds of organisms (e.g. the Kyoto Encyclopedia of Genes and Genomes (KEGG) or the BioCyc database) where relations between genes, enzymes, reactions and chemical compounds are available. The knowledge that we have of these relations is however incomplete (most annotation efforts fail to assign functions to 40-60% of the sequences [Y. Pouliot, P.D. Karp, BMC Bioinformatics 8:244, 2007]) and is affected by uncertainty (wrong EC number assignment, incomplete annotation (e.g. only one function of a multidomain protein) or nonspecific assignment (e.g. to a protein family)). Systems capable to perform automatic curation of these databases and capable to suggest pathway-holes fillings are therefore needed. For this purpose one can make use of information on related organisms and use evidence based not exclusively on homology searches, but also on genomic and/or functional context. This raises the problem of how to integrate heterogeneous and uncertain sources of information in a principled way. Although systems for reconstructing pathways from relevant gene sets (KAAS [Y.Mariya et al. Nucleic Acids Res, 35:W182-W185, 2007]) and filling pathway-holes (PathoLogic [M.L. Green, P.D. Karp, BMC Bioinformatics 5:76, 2004]) are known in literature, they do not offer sufficient flexibility when new additional sources of information become available or, more importantly, in case one needs to change the set of queries involved in the solution of a specific task.

Here we propose an approach that satisfies these flexibility requirements: we represent metabolic networks in a probabilistic logical framework. In this way background knowledge affected by uncertainty can be easily included, and we can obtain an answer to several key questions performing probabilistic inference in a principled manner. More specifically, we use ProbLog [A. Kimmig et al., LNCS vol.5366, 2009], a simple yet powerful extension of the logic programming language Prolog with independent random variables in the form of *probabilistic facts*. In contrast to propositional graphical models (such as Bayesian Networks), connections between random variables in ProbLog can be specified on the first order level, thus avoiding the need of explicitly grounding all information a priori, achieving therefore a higher abstraction and flexibility in the queries specification.

In this work we start to investigate three fundamental problems concerning metabolic networks: 1) *network prediction*, that is, the reconstruction of a metabolic network when given information on the set of the relevant organism's genes and their probable enzymatic functions 2) *link prediction*, i.e. the correction of the link strenght between a gene and an enzyme, and 3) *node prediction*, that is, whether the existence of a certain enzyme (and hence of an unknown gene) has to be hypothesized in order to maintain the contiguity of highly probable pathways. We report encouraging empirical results on the KEGG dataset. Here we measure the fraction of correctly retrieved links and nodes and show that the probabilistic approach outperforms a purely deterministic solution to the same problems.